Data Scientists Are Freaks of Nurture but Products of Nature

ESIP Summer Data Science and Analytics Technical Session

July 14, 2015, Asilomar CA

Peter Fox (RPI and WHOI/AOP&E) pfox@cs.rpi.edu, @taswegian
Tetherless World Constellation, http://tw.rpi.edu #twcrpi
Earth and Environmental Science, Computer Science, Cognitive Science, and IT and Web Science

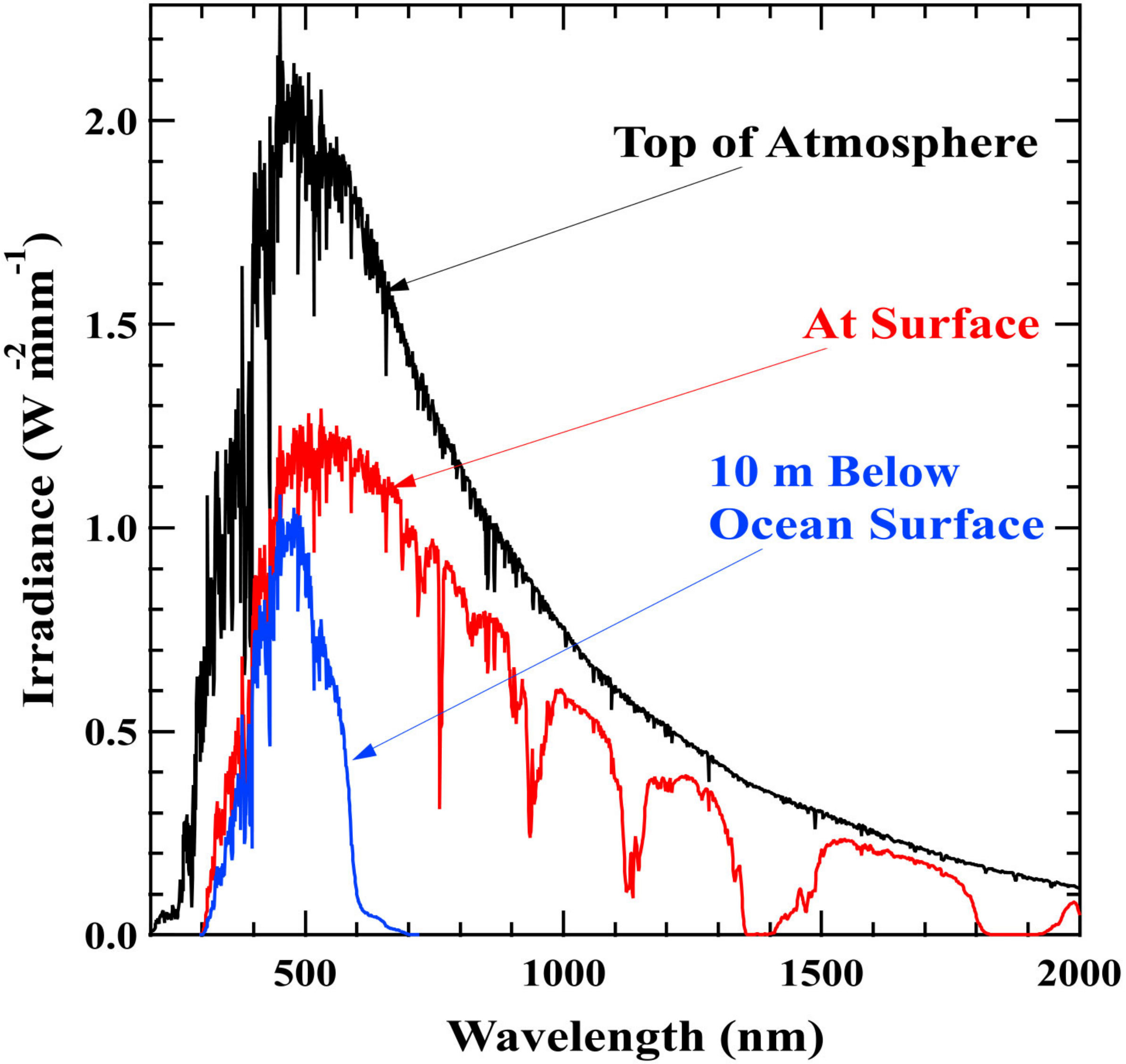
Why do we (I) care about the Sun?

- The Sun's radiation is the single largest external input to the the Earth's atmosphere and thus the Earth system.
- Add, it varies – in time and wavelength
- Also, for a long time – Solar Energetic Particles and the near Earth environment (and more recently the effect on clouds?)
- Observations commenced ~ 1940's, with a resurgence in the late 1970's
- Two quantities of scientific interest
  - Total Solar irradiance – TSI in Wm^{-2} (adjusted to 1AU)
  - Solar Spectral Irradiance – SSI in $Wm^{-2}m^{-1}$or $Wm^{-2}nm^{-1}$
- Measure, model, understand -> construct, predict

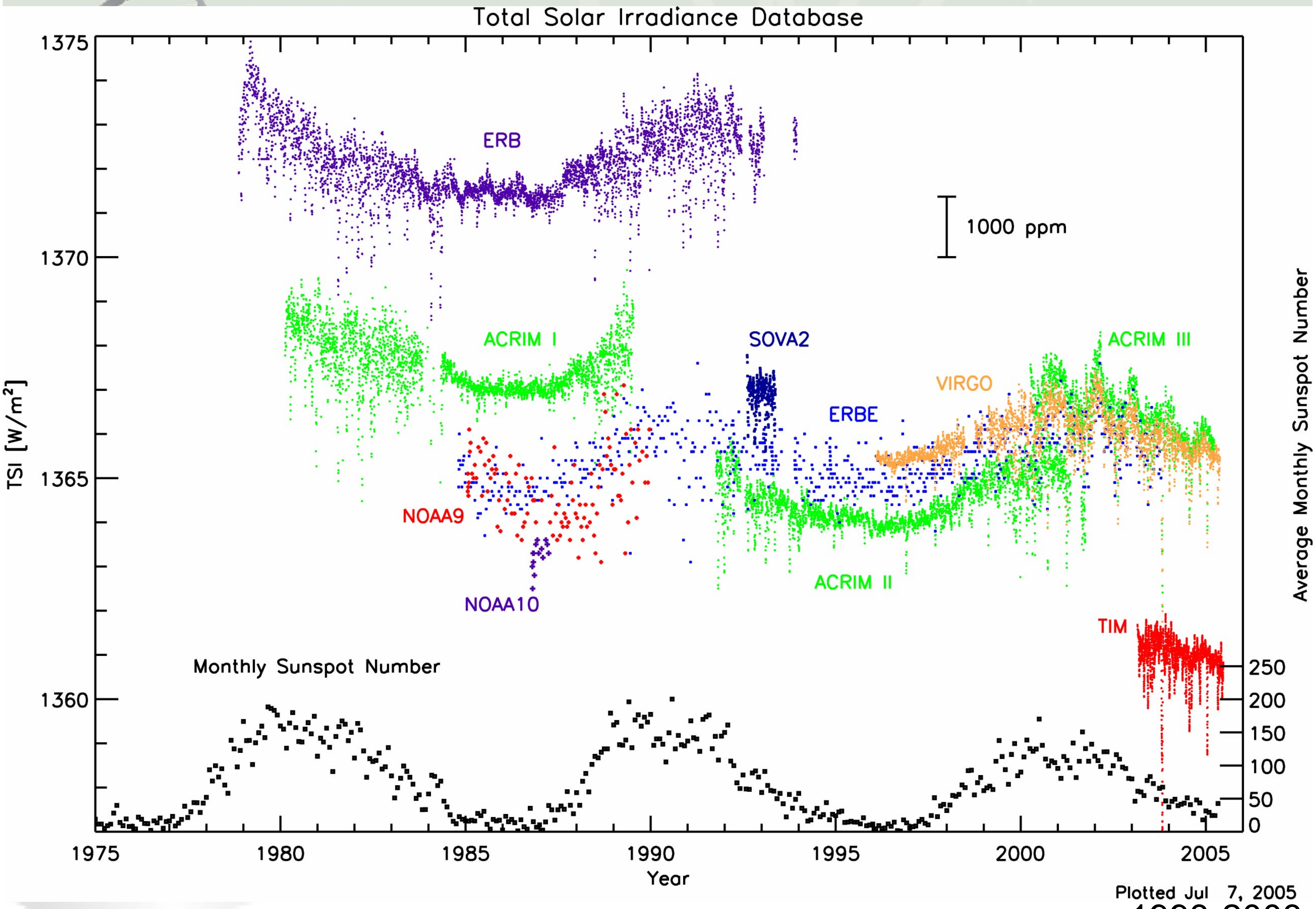
pfox@yale 1986-1991 and pfox@ncar 1991-2007

1993-2003

Solar radiation as a function of altitude

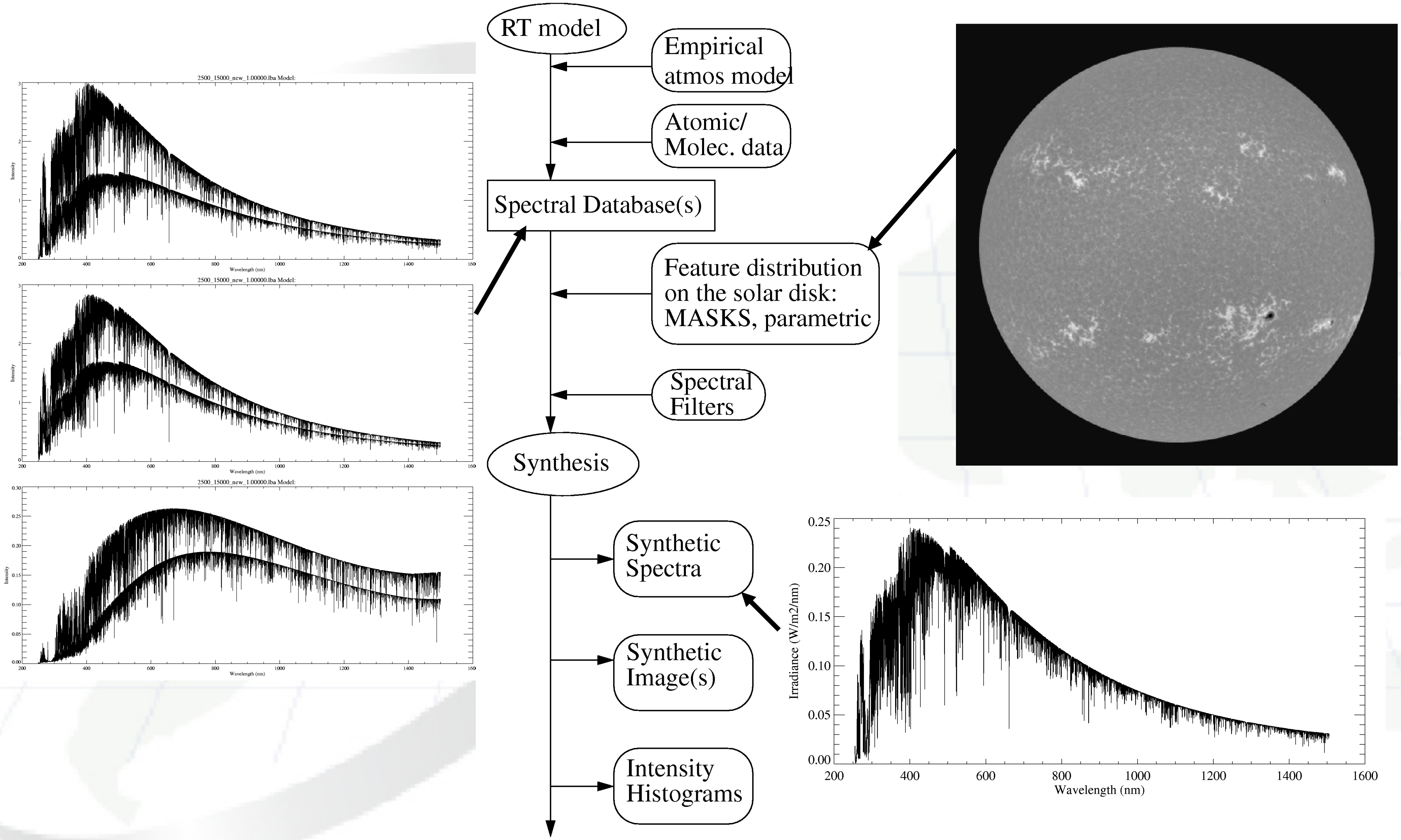


1993-2003


Total Solar Irradiance Database
TSI [W/m²]
1375
1370
1365
1360
Average Monthly Sunspot Number
250
200
150
100
50
0
1975
1980
1985
1990
1995
2000
2005
Year
ERB
1000 ppm
ACRIM I
SOVA2
ACRIM III
VIRGO
ERBE
NOAA9
NOAA10
ACRIM II
TIM
Monthly Sunspot Number
Plotted Jul 7, 2005

1993-2003

Spectral synthesis components and flow



RT model

Empirical atmos model

Atomic/ Molec. data

Spectral Database(s)

Feature distribution on the solar disk: MASKS, parametric

Spectral Filters

Synthesis

Synthetic Spectra

Synthetic Image(s)

Intensity Histograms

1993-2003

Summary of Results

- First comprehensive ‘database’ of:
  - Empirical models of the thermodynamic structure of the solar atmosphere suitable for different solar magnetic activity levels
- First comprehensive (70 component) synthetic spectral irradiance database in absolute units
  - 10 disk angles, 7 models, far ultra- violet to far infrared, multi-resolution
  - ~724 GB (in 1995)
- Strong validation in ultraviolet, visible, lines, infrared
  - Correct center to limb prediction for red-band irradiances
  - Found 30-45% network contribution to Ly-α irradiance
- Several comparisons led to improvements in the atomic parameters
- Led to choice of PICARD (new satellite) filter wavelengths

1993-2003

Which brings us to DATA SCIENCE

- Drum roll…..
- Some dirty secrets
- And some … universal truths…

Needs (:== mantra)

Scientists should be able to access a global, distributed knowledge base of scientific data that:

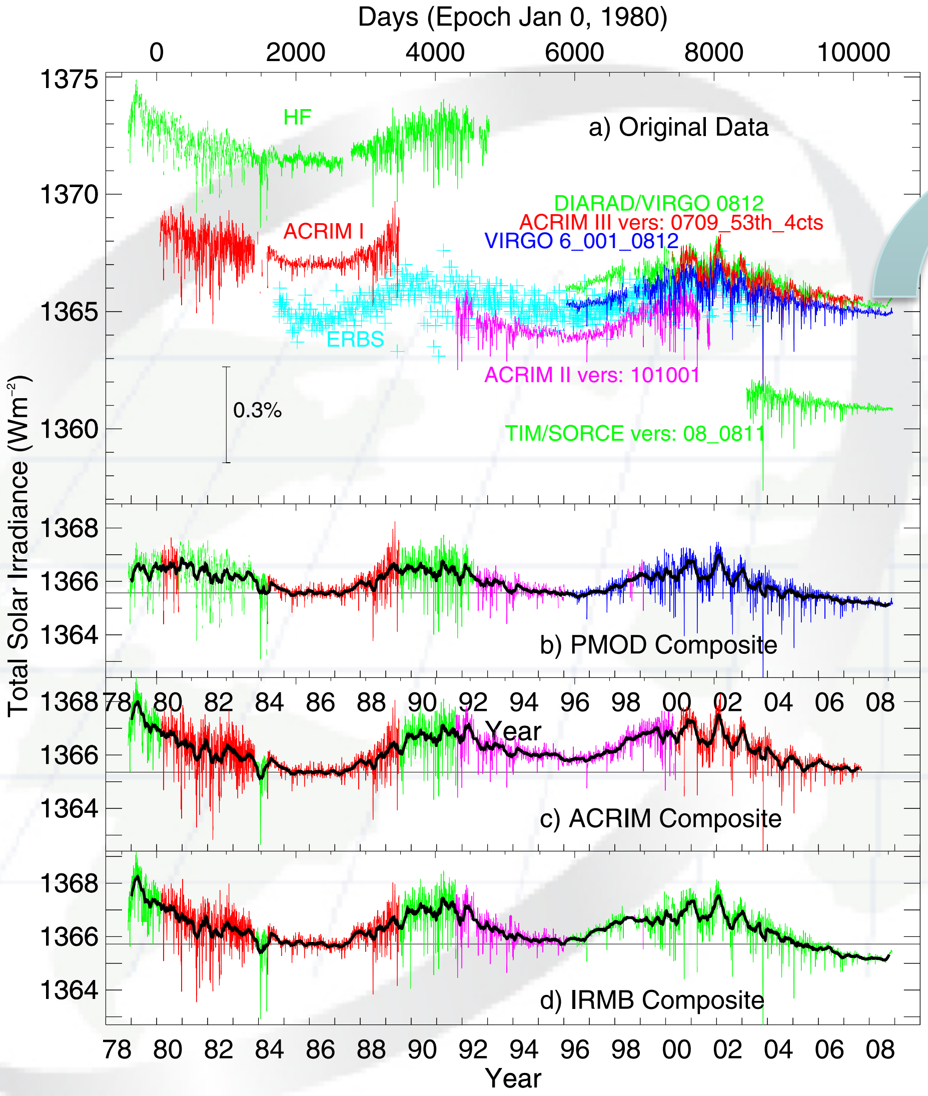
- appears to be integrated
- appears to be locally available

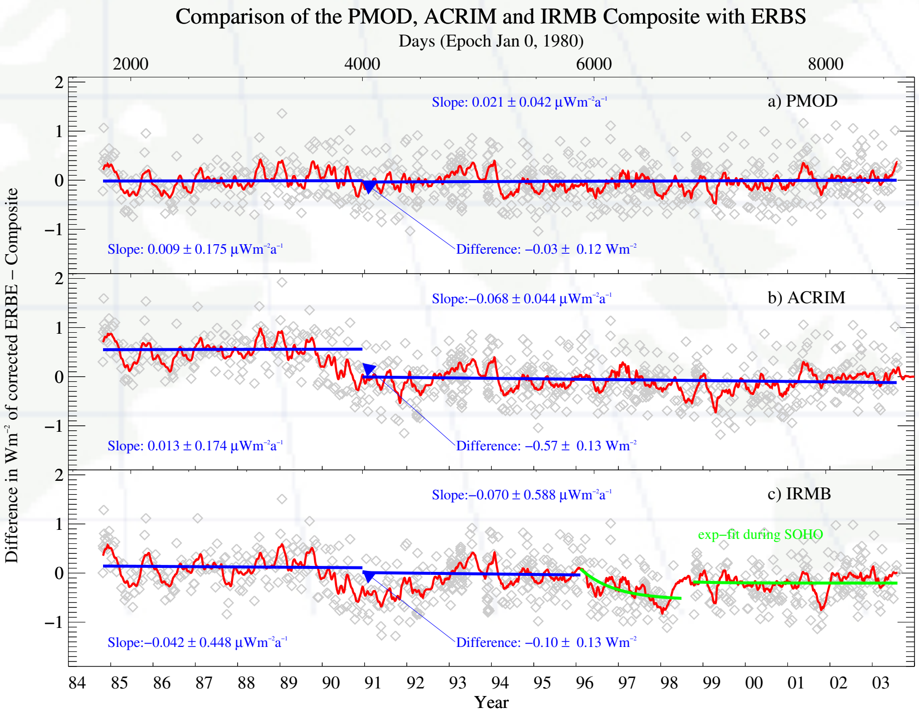
But… data is obtained by multiple means (models and instruments), using various protocols, in differing vocabularies, using (sometimes unstated) assumptions, with inconsistent (or non-existent) meta-data. It may be inconsistent, incomplete, evolving, and distributed. And created in a manner to facilitate its generation NOT its use.

And… there exist(ed) significant levels of semantic heterogeneity, large-scale data, complex data types, legacy systems, inflexible and unsustainable implementation technology

1997

Back to the TSI time series…

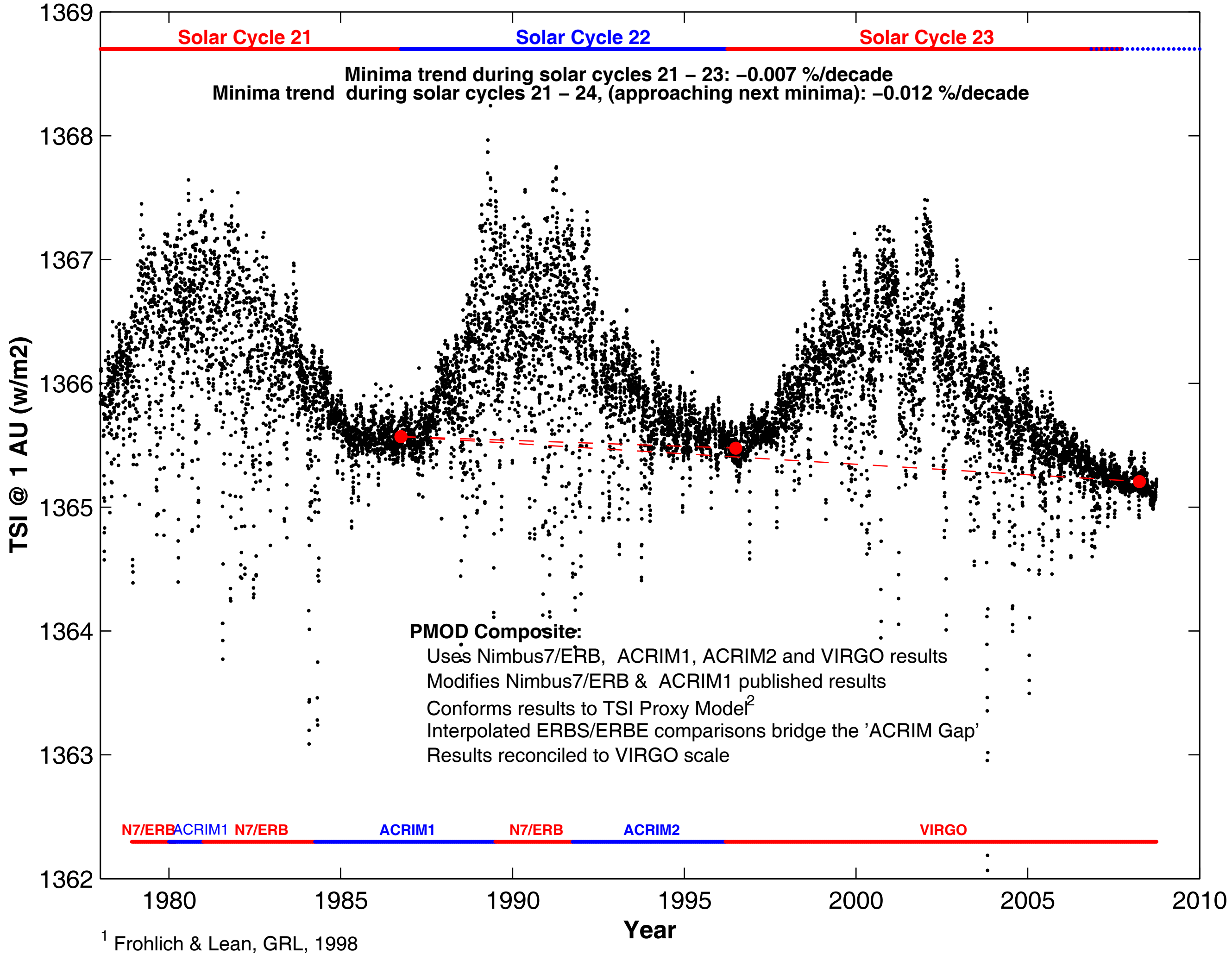



Comparison of the PMOD, ACRIM and IRMB Composite with ERBS

1993-2003

One composite, one assumption

PMOD Composite TSI Time Series (Daily Means)[1]



Minima trend during solar cycles 21 – 23: –0.007 %/decade
Minima trend during solar cycles 21 – 24, (approaching next minima): –0.012 %/decade

PMOD Composite:
- Uses Nimbus7/ERB, ACRIM1, ACRIM2 and VIRGO results
- Modifies Nimbus7/ERB & ACRIM1 published results
- Conforms results to TSI Proxy Model[2]
- Interpolated ERBS/ERBE comparisons bridge the 'ACRIM Gap'
- Results reconciled to VIRGO scale

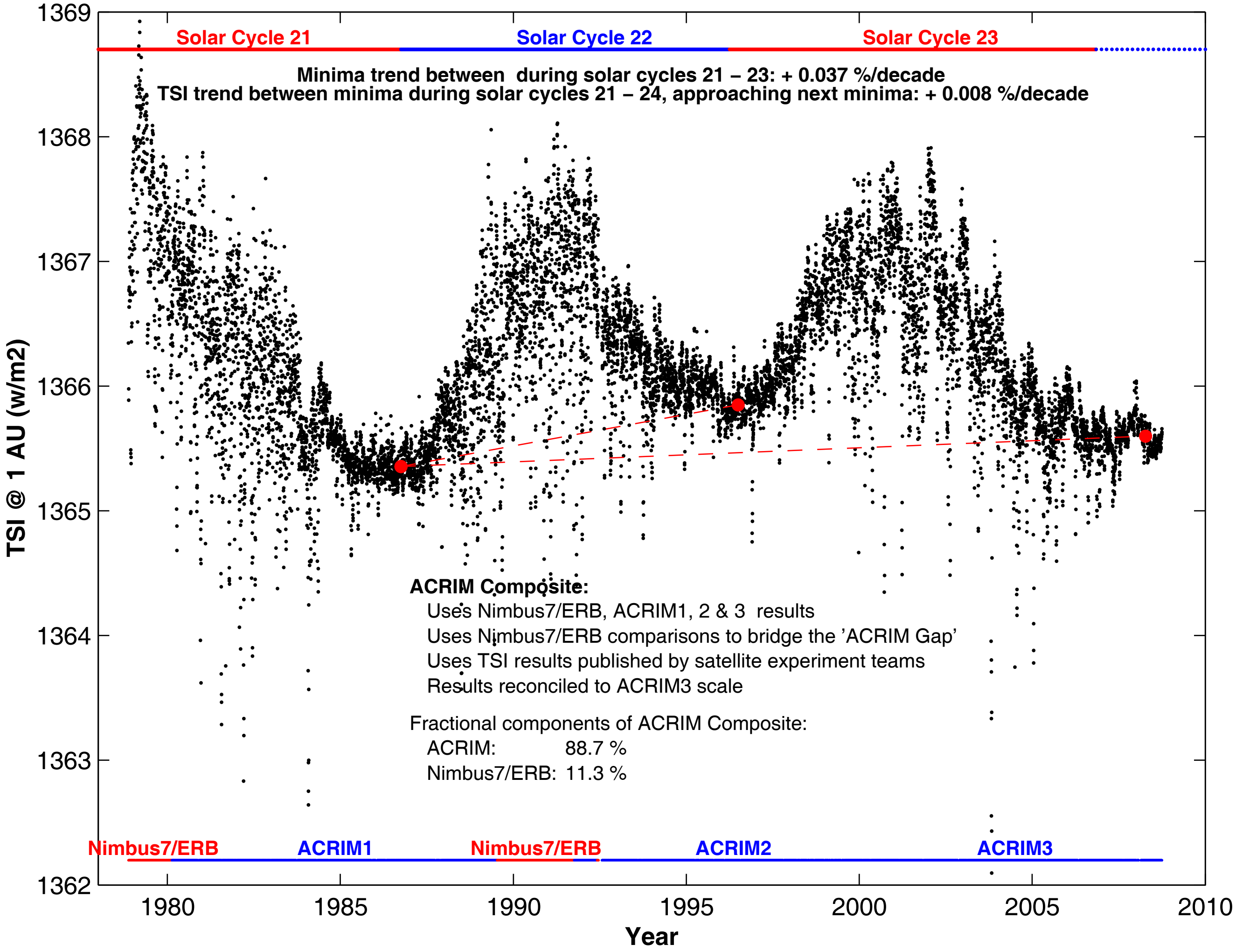
RC Willson, earth_obs_fig27 11/22/2008

[1] Frohlich & Lean, GRL, 1998

[2] Lean, Beer & Bradley GRL, 1995

Another composite, different assumption

ACRIM Composite TSI Time Series (Daily Means)[1]



Minima trend between during solar cycles 21 – 23: + 0.037 %/decade
TSI trend between minima during solar cycles 21 – 24, approaching next minima: + 0.008 %/decade

ACRIM Composite:
Uses Nimbus7/ERB, ACRIM1, 2 & 3 results
Uses Nimbus7/ERB comparisons to bridge the 'ACRIM Gap'
Uses TSI results published by satellite experiment teams
Results reconciled to ACRIM3 scale

Fractional components of ACRIM Composite:

ACRIM:	88.7 %
Nimbus7/ERB:	11.3 %

[1] Willson & Mordvinov, GRL, 2003 **RC Willson, earth_obs_fig26 11/23/2008**

Data pipelines: we have problems

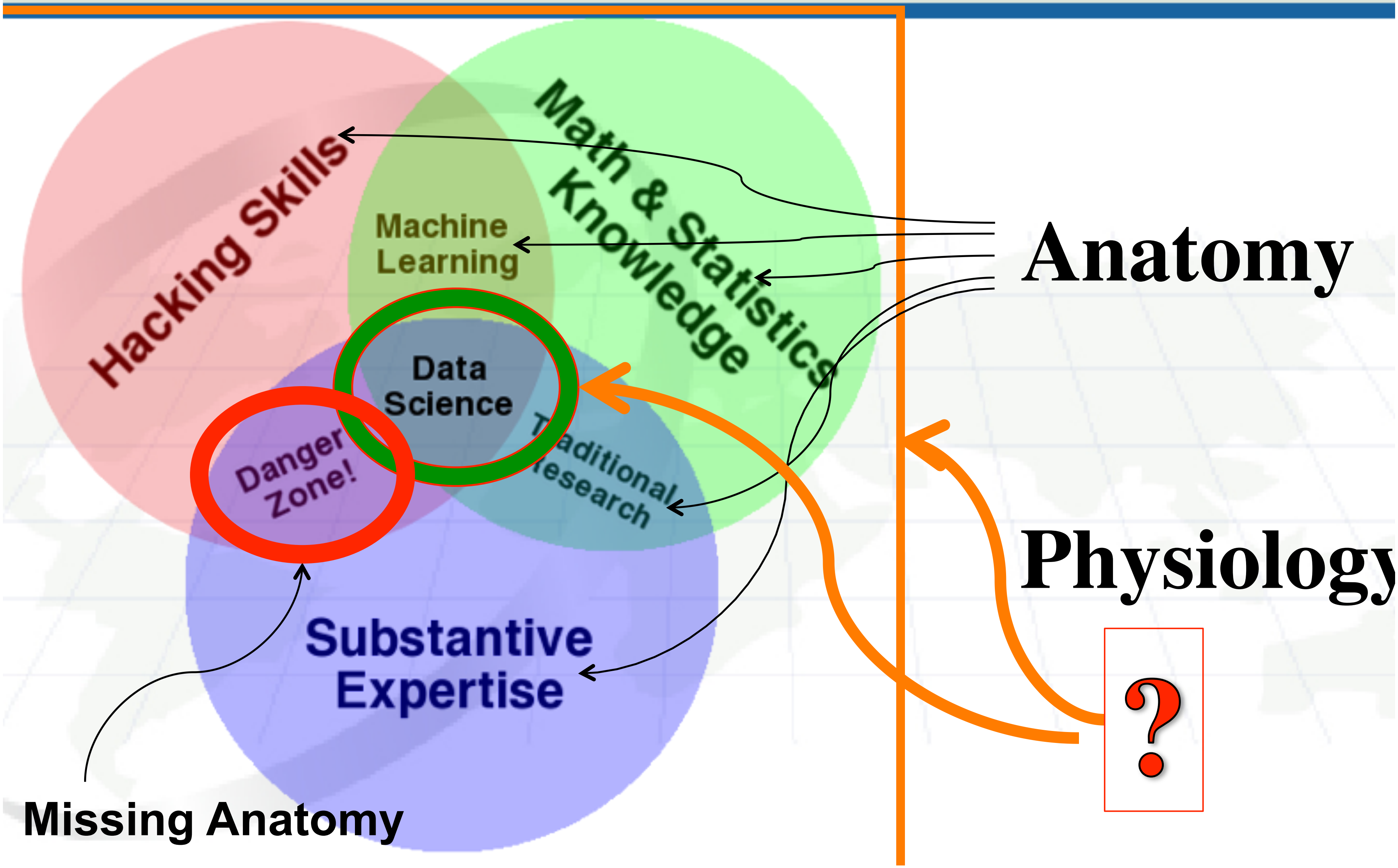
- *Data is coming in faster, in greater volumes and forms and outstripping our ability to perform adequate quality control*
- *Data is being used in new ways and we frequently do not have sufficient information on what happened to the data along the processing stages to determine if it is suitable for a use we did not envision*
- *We often fail to capture, represent and propagate manually generated information that need to go with the data flows*
- *Each time we develop a new instrument, we develop a new data ingest procedure and collect different metadata and organize it differently. It is then hard to use with previous projects*
- *The task of event determination and feature classification is onerous and we don't do it until after we get the data*
- *And now much of the data is on the Internet/Web (good or bad?)*

Metaphor

- **Anatomy** study of the structure and relationship between *body* parts

- **Physiology** is the study of the function of body parts and the body as a whole.

Overused Venn diagram of the intersection of skills needed for Data Science (Drew Conway)



Anatomy

Physiology

?

Missing Anatomy

Data Science

- ✧ **Anatomy (as an individual)**
  - ✧ **Data Life Cycle – Acquisition, Curation and Preservation**
  - ✧ **Data Management and Products**
  - ✧ **Forms of Analysis, Errors and Uncertainty**
  - ✧ **Technical tools and standards**

Data Science

- ✧ **Physiology (in a group)**
  - ✧ **Definition of Science Hypotheses, Guiding Questions**
  - ✧ **Finding and Integrating Datasets**
  - ✧ **Presenting Analyses and Viz.**
  - ✧ **Presenting Conclusions**

Data Analytics

- ✧ **Anatomy (individual)**
  - ✧ Intermediate Skill in parametric and non-parametric statistics
  - ✧ Application of a broad spectrum of Data Mining and Machine Learning Algorithms
  - ✧ Ability to cross-validate and optimize models
  - ✧ Application to specific datasets

Data Analytics

- **Physiology (term project)**
  - **Definition of Science Hypotheses, with Prediction/ Prescription Goal**
  - **Cleaning and Preparing Datasets**
  - **Validating and Verifying Models**
  - **Presenting Ideas and Results**

Call to Action – Data Science

- Data Science across the curriculum
  - Same as “Calculus”
  - And … Intro to Statistics
- Data Management is Second Nature
  - Like operating an instrument
  - Openness/ sharing is the natural state
- As-a-whole, the Data Scientist works collaboratively and is recognized and rewarded by peers and organizations

Call to Action – Data Analytics

- Institutions to provide reliable, high-functionality data infrastructures that facilitate analytics
- Provision of intermediate to advanced Statistics to undergraduates and early graduate students
- Well-curted datasets are made widely available along with developed models and validation statistics
- All results are under continuous scrutiny, are traceable and verifiable

pfox@rpi = 6-7 years in (2008-) … what made "me"?

- Science and interdisciplinary from the start!
  - Not a question of: do we train scientists to be technical/data people, or do we train technical people to learn the science
  - It's a skill/ course level approach that is needed
- Teach methodology and principles over technology
- Data science is a *skill*, and natural like using instruments, writing/using codes
- Team/ collaboration aspects are key
- Foundations and theory must be taught

See also…

- http://tw.rpi.edu/media/latest/AGU2014-ED31E-3455_Fox.pptx
- “Training Students to Extract Value from Big Data: Summary of a Workshop”
  - http://sites.nationalacademies.org/DEPS/BMSA/DEPS_087192
  - http://www.nap.edu/catalog/18981/training-students-to-extract-value-from-big-data-summary-of